Machine Learning for Natural Language Processing

Arieda Muço

Central European University

Information

- In my research, I deal with tons of data and (lots of) text data. That's why this course.
- Introduce yourself. What are your expectations? Why are you here? What kind of text/data you are currently using or plan to use?

Plan for this course

- Supervised Learning
- Unsupervised Learning
- Applications on Text Data
- Implementation of algorithms with Python

Grading

Final assessment will consist of the following:

- **In classroom participation** (20% of final grade)
  - Quizzes will be given in class but some will have a longer deadline. We will discuss them in class. Collaboration is encouraged
- **2 Problem Sets** (40% of final grade) – Submission in groups
- **Individual Project** (40% of final grade)
- **Deadlines** Past deadlines submissions are accepted but you'll get -1 point for each hour past the deadline. So, if you submit 5 hours past deadline, you'll be deducted 5 points.

Organization

- Slack will be our communication tool for this course
  - Post questions and answers in respective channels
  - Keep a close eye on channels on quizzes and assignments
  - Make sure you reply in thread when needed
- I strongly encourage peer learning. Feel free to post in the Slack channel if you think some information is of common interest
- Questions: You will get a reply. However, questions about homework should be received by us 48 hours before a deadline (no response otherwise)

Rules

- Make sure you do the readings assigned and go through slides and notebooks (you will be quizzed on material we discuss in class)
- Try to type code along with me
- Ask questions and feel free to use AI/Google
  - Don't feel bad about this especially for the programming part of the course.
  - Important to know how to read error messages
    - or google them
  - ChatGPT and Stack Overflow(?) are a programmer's best friend

Recommended Material

- Python
  - Codecademy is the place to start
  - Automate the Boring Stuff with Python and The Real Python are great sources
- Machine Learning
  - An Introduction to Statistical Learning (ISL) by Gareth, Witten, Hastie and Tibshirani
  - The Elements of Statistical Learning (ESL) by Hastie, Tibshirani, Friedman
  - Statistical Learning with Sparsity (SLS) by Hastie, Tibshirani, Wainwright
  - Introduction to Machine Learning with Python: A Guide for Data Scientists (IMLP) by Sarah Guido, and Andreas Muller
- Text Analysis
  - Introduction to Information Retrieval by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze
  - Speech and Language Processing by Dan Jurafsky and James H. Martin

Academic Papers in Economics

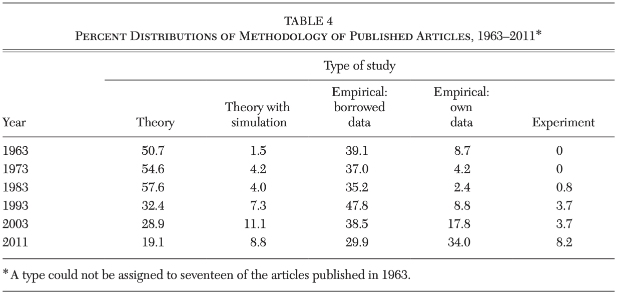

TABLE 4
Percent Distributions of Methodology of Published Articles, 1963–2011*

	Type of study				
Year	Theory	Theory with simulation	Empirical: borrowed data	Empirical: own data	Experiment
1963	50.7	1.5	39.1	8.7	0
1973	54.6	4.2	37.0	4.2	0
1983	57.6	4.0	35.2	2.4	0.8
1993	32.4	7.3	47.8	8.8	3.7
2003	28.9	11.1	38.5	17.8	3.7
2011	19.1	8.8	29.9	34.0	8.2

*A type could not be assigned to seventeen of the articles published in 1963.

Hammermesh (2013)

Background

- Old data, structured and small: (gdp, population, investment)
- New data, less structure and larger (scraped data, consumer search patterns, social networks, texts, ?)
- New methods needed: data collection/management, workflow/collaboration, description/analysis

Causal Inference and Machine Learning

- Causal Inference
    - Focus on one/few coefficients of interest (causal effect)
    - Use one main specification, show robustness to alternative specification and placebo tests
    - Model rarely evaluated (when pure inference we focus on in-sample-properties, mostly R^2)
- Machine Learning (ML)
    - Focus on prediction (and description)
    - Use data-driven model selection to have best prediction (treated as a black box)
    - Model is evaluated out-of-sample (e.g. cross-validation)

Use ML to identify the most meaningful predictive variables (i.e Lasso and Ridge), dimensionality reduction, generate outcome of interest Y, or/and main variable of interest X

Linguistic differences

	Econometrics	Machine Learning
Y	Outcome	Target
X	Independent Variables	Features

Note that Scikit-learn and IMLP refer to observations as "Samples". Don't be confused!

Supervised vs Unsupervised Learning

- Supervised Learning: Y, the target, is available. Labeled data
  - Regression: Y is continuous
  - Classification: Y is categorical (binary or multi-class – ordered or not ordered)
- Unsupervised Learning: Y is not available
  - Exploratory data analysis and can be useful as a pre-processing step for supervised learning

Other types of learning

- Deep Learning
- Semi-Supervised
- Active Learning
- Forecasting

Know Your Task

- Each algorithm is different in terms of what kind of data and what problem setting it works best for. When building an algorithm ask:
  - What question(s) am I trying to answer? Do I think the data collected can answer that question?
  - What is the best way to phrase my question(s) as a machine learning problem?
  - Have I collected enough data to represent the problem I want to solve?
  - What features of the data did I extract, and will these enable the right predictions?
  - How will I measure success in my application?
  - How will the machine learning solution will help my project?

Know Your Data

- The most important task when working with data is knowing your data
  - All data related work
  - Extract features only if you know your data well enough. We are going to talk about best practices throughout this course

A bit about Python

- Programming language intended for general-purpose high-level language
- Web development, scientific and numeric education, desktop graphical user interface, software development
- Free and open source
- You can do everything that you can do in a programming language
- Big community (Google, Youtube, Nasa...)
- High readability (more than R or C)
- Python was first released in early 1980
    - Python 2 in 2000 and Python 3 in 2008

Black Holes and Python

Purpose of the course

- Machine Learning and Text Analysis, as well as programming in Python, are (mildly put) very broad topics, and we will not be able to cover many(!) things
- Build foundations such that in the future you get confidence in starting to dig deeper into these topics
- Strong focus towards applications and real-life problems